UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SCANSOFT, INC., Plaintiff v. VOICE SIGNAL TECHNOLOGIES, INC., LAURENCE S. GILLICK, ROBERT S. ROTH, JONATHAN P. YAMRON, and MANFRED G. GRABHERR, Defendants	C.A. No. 04-10353-PBS

DECLARATION OF BRUCE BALENTINE
CONCERNING VOICE ACTIVATED DIALING TECHNOLOGY

I, Bruce Balentine, declare as follows.

1. I am a consultant in the speech recognition industry, in which I have worked for the past twenty years. I believe that, over the years, I have gained expertise in the design and development of commercial speech recognition systems, such as Voice Activated Dialing ("VAD") products, which are at issue in this litigation. For example, I have designed and developed many commercial applications of speech recognition systems for Fortune 500 companies. I also am the inventor or co-inventor of several patents in this field. I have authored a best-selling book and other publications on the design of speech recognition applications and have lectured widely on this subject at speech recognition industry workshops, conferences, seminars, and trade shows. Accordingly, when ScanSoft's counsel asked me to provide a tutorial on speech recognition technology, I felt comfortable doing so. What follows below is my discussion of the art of speech recognition systems, particularly as it relates to United States Patent No. 6,501,966 ("the '966 patent").

CREDENTIALS

2. As prefaced above, I specialize in the design and development of commercial applications for speech recognition systems, such as those used in telecommunications and in automated call centers. For example, I have designed several voice-activated dialing ("VAD") applications for mobile cell phone networks. One of the VAD applications that I designed, the Voice Dialer 2060 product, earned the Home Automation Association's Mark of Excellence 1995 User Friendly Hardware Award. I have also designed speech-enabled voicemail systems and a variety of other applications relying on speech recognition.

3. I am currently Chief Scientist and Executive Vice President of Enterprise Integration Group ("EIG"), a consulting, research, and engineering firm specializing in the design of Interactive Voice Response ("IVR") systems, which are commonly used to automate customer service lines or "call centers." While at EIG, I have developed IVR systems for banks and other financial services companies, telephone companies, and a variety of Fortune 500 corporations.

4. I have focused my work in the speech recognition field on the study of "human factors." The study of "human factors" reveals how people actually operate in everyday situations--in this case, how they speak and use a phone or computer. I use human factors research to improve the way that speech recognition systems interact with users, making the experience more natural, user-friendly, and less error prone. Indeed, the most important benefit of human factors analysis is that it helps prevent or reduce errors in speech recognition systems. By anticipating how people use speech recognition systems in everyday situations, we can design systems that are more robust and better able to catch and correct errors when they occur. (A typical error in a speech recognition product, such as a VAD application, could be, for example,

when the system mistakes the command "Call Home" for the command "Call Office.") We have found that it is more effective to reduce such errors at the user interface level rather than only at the speech recognition "engine" level (that is, the software programs and mechanisms that are used to recognize sounds).

5. I did similar design and development work as a speech recognition consultant for a sole proprietorship that I founded in 1995 called Same Page Design Group. I merged Same Page Design with Enterprise Integration Group in July 2000. While at Same Page Design, I designed and developed VAD applications for the telecommunications industry.

6. From 1986 through July 1995, I was Director of Product Marketing and Manager of Applications Development for Scott Instruments, a developer of speech recognition systems for the telecommunications industry. Scott also developed speech recognition products used in voice-based training systems for the U.S. military, used in multimedia desktop applications (i.e., for personal computers), and even for toys.

7. Scott Instruments merged with Voice Control Systems in 1994, and that combined entity was eventually purchased by Philips Speech Processing. I stayed at this merged entity for one year and left in July 1995, when I founded Same Page Design Group. I note that Voice Control Systems was the employer of the three inventors of the '966 patent. I worked with those inventors--Bern Bareis, Pete Foster, and Tom Schalk--for a short time after the merger of Scott Instruments and Voice Control Systems, but I did not have any involvement with the patent or the inventions that are the subject of that patent.

8. Before joining Scott Instruments, I worked for several years in the personal computer industry as an engineer and applications consultant. In particular, I helped develop

home and office automation systems. I also consulted to a telecommunications software company on high-speed data transfer applications in personal computer networks.

9. As mentioned above, I have authored books and articles in the field of speech recognition, have frequently lectured at industry events, and have otherwise devoted my professional career to improving applications of speech recognition systems, as more fully detailed in my C.V., a copy of which is attached as Exhibit A.

10. Finally, I am an inventor or co-inventor of several patents in the field of speech recognition, including, for example, United States Patent No. 5,025,471, entitled "Method and Apparatus for Extracting Information-Bearing Portions of a Signal for Recognizing Varying Instances of Similar Patterns," issued on June 18, 1991. This patent, broadly speaking, concerns methods and equipment for improving "speaker independent" speech recognition. A copy of this patent is attached as Exhibit B. Later in this declaration I will discuss the concepts of "speaker independent" and "speaker dependent" speech recognition systems, both of which come into play in various embodiments of the speech recognition applications claimed in the '966 patent.

MY ASSIGNMENT

11. ScanSoft's lawyers have asked me to comment on the '966 patent, with a particular view to explaining the technology it discloses and the technological and commercial background against which the patent was developed. I have also been asked to interpret technical jargon, if any, used in the patent so that those unfamiliar with speech recognition technology can better understand what the patent is talking about. Finally, I have been asked to comment on how "one of ordinary skill in the art" would have read and understood the patent.

12. Accordingly, I began my assignment by reading the '966 patent, its prosecution history, related patents in its family (*e.g.*, United States Patent No. 5,297,183, which is the first

patent that issued from the original application), and various prior art patents and publications. I note that the '966 patent is based on an application filed in the Patent Office in November 2000 but that the application is a continuation of a chain of applications dating back to April 1992. The prior applications all resulted in patents, and I have reviewed these and compared them with the '966 patent. All of these patents share the same specification but claim different components or variations of a speech recognition system. I understand that ScanSoft accuses Voice Signal Technologies ("VST") of infringing Claims 1-6 of the '966 patent, so I have read these claims in detail.

13. The '966 patent relates to certain speech recognition systems adapted for use in mobile communications systems, and more particularly for systems that recognize "spoken commands and for the directing of telephone calls based on spoken commands." '966 patent, Col. 1, *ll.* 15-20. In other words, the '966 patent concerns the use of a speech recognition system to allow telephone users (and preferably mobile cell phone users) to dial phones through voice activation (what is known in the industry as "VAD," or Voice Activated Dialing). VAD is a process by which the cell phone user's spoken commands substitute for punching numbers on the telephone keypad.

14. I am told by ScanSoft's counsel that the relevant standard for understanding a patent is how "one of ordinary skill in the art" would have understood it at the time the patent application was filed. I have learned from ScanSoft's counsel that a common definition of "one of ordinary skill in the art" is an average worker in the field--such as an average engineer, mechanic, or chemist, etc.--who is not an inventor in his or her own right but who can nonetheless follow the teachings of the patent to build an actual embodiment of the patented inventions. In this case, the relevant time frame is 1992, when the original application in the

chain was filed. I feel qualified to comment on how one of ordinary skill would have understood the '966 patent as of April 1992 because I was working in the field at that time--and indeed, for many years prior--designing voice activated dialing and other speech recognition products for the telecommunications, office automation, and call center industries. Indeed, by 1992, I had already been awarded a patent in the field and had already authored a number of published books and articles, as cataloged in my C.V.

15. I believe that one of ordinary skill in the art, circa 1992, would be a person working in the field of speech recognition products for several years and having an education or work experience in software development and/or software product management. This person might be someone working for a speech recognition products company--such as Scott Instruments or Voice Control Systems or one of their competitors at the time--or, perhaps, someone working for a customer of a speech recognition company, such as a telecommunications company or cell phone manufacturer. This "ordinary person" might also have at least some background in or knowledge of telephony--whether from the technical perspective or at least from a marketing or product management perspective.

TUTORIAL ON SPEECH RECOGNITION

16. The goal of speech recognition systems is to allow users to control or direct computer functions with their voices. A user's voice, in essence, becomes a computer keyboard or telephone keypad. Thus, a user may call a speech-driven customer service hotline to activate a credit card, to check a bank account balance, or to check on the status of an airline flight, without the need for a computer keyboard and mouse. Likewise, a user's voice may be used to dial a telephone number without having to look at and punch the keys on a touch tone phone. This application is particularly useful for "hands-free" use of a mobile cell phone while driving.

Indeed, hands-free use of car phones was one of the primary motivations for developing VAD systems during the mid- to late 1980s, when the mobile phone industry was emerging.

17. How does this speech recognition work? How is it deployed in telecommunications networks? A basic understanding of speech recognition systems would assist one to understand better the inventions claimed in the '966 patent. I present below a very simplified explanation of speech recognition systems.

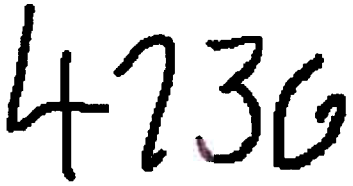
18. In the abstract, speech recognition attempts to mimic what even toddlers do easily and subconsciously--recognize and understand spoken words. Training computers to recognize speech has been underway since at least the 1950s but has proven difficult to perfect. Speech recognition is still marked by a high rate of error. For example, take two words commonly used in VAD applications, "call" and "dial." People hear and recognize these terms as two completely different words. But computers might recognize them as the same. Over the years, software algorithms have been developed that use statistical models or other schemes to predict with increasing confidence what word was actually spoken. So in a sense, computerized speech recognition is not an exact science but is rather a set of predictions of the likelihood that a particular word or number has been spoken. Sometimes the prediction is wrong. Reducing or preventing errors in speech recognition is thus critical to improving the process.

19. A common misconception is that speech recognition systems actually "hear" and "understand" words spoken by a user. In reality, and at a very basic level, computerized speech recognition works by characterizing sounds spoken by the user into acoustic properties, such as frequency and amplitude, and assigning numeric values to these properties. These numeric values are then analyzed and "scored" using complex word and language models and statistical

analysis to determine the words spoken. By "scored," I mean determining the probability that the word spoken was, in fact, "call" rather than "dial" based on the numeric values.

20. That is, because speech varies from speaker to speaker depending on factors like accent, geography, and even emotional states, statistical analysis must be performed to predict the probability that a word like "call" is not "dial" or some other word. Speech recognition thus relies on statistical models to confirm the accuracy of a hypothesis that the sounds equal a particular word. Of course, this recognition process is done at very high speeds by software and logic circuits of the computer processor, so to the user, the system appears to hear and understand the spoken words. What happens below the surface is actually more complex.

21. The basic problem limiting speech recognition systems is error rate caused by these variations in speech from speaker to speaker. Human beings are adept at recognizing a word that was not fully heard, or is shrouded in a thick accent, based on context and other cues. Speech recognition systems, however, have difficulty distinguishing words based on context. Take, for example, the following representation of a handwritten number:



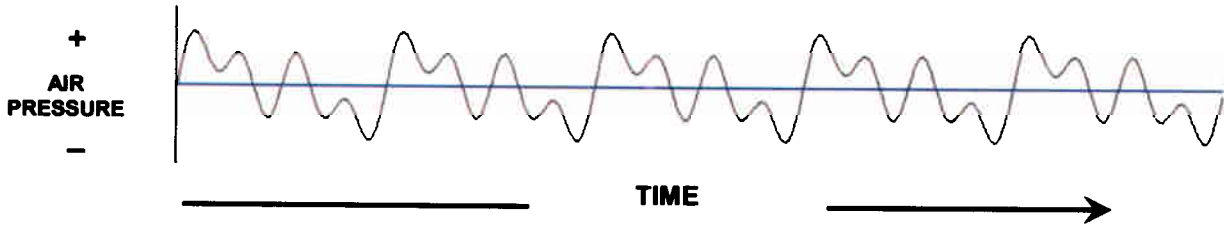
22. A human being might recognize the first digit as 4, the next digit as either 2 or 7, and the last digit as either 0 or 6. How confident are you that the second number is, say, 2? This is the process that a speech recognition system must go through to distinguish words and numbers spoken by the user--especially in a system used in a cell phone network, in which the words spoken by the user might be distorted by the hum of the engine, the car radio, or other

noise. The system uses statistical analysis or other methods to determine the probability that the second digit is, in fact, 2 and not 7. I will discuss this process in a bit more detail below.

A. Basic Units of Speech

23. To understand how a speech recognition system works, one must first step back and understand what speech is. In general, speech is a series of words or utterances spoken in an order (i.e., a "grammar") understood by the listener. Words, in turn, are composed of "phonemes." A phoneme is the simplest linguistic unit. The "c" or "t" in "cat" is a phoneme. Speech recognition works by detecting the phonetic patterns of speech and attempting to associate them with known words programmed into the system's vocabulary. Early recognizers detected these patterns as embedded in whole words. Modern systems operate at the level of the phoneme.

24. On an even more basic level, phonemes are composed of sounds. Sound is nothing more than rapid increases and decreases in air pressure against one's ear drum. Sound (*i.e.*, the variations in air pressure) is represented as a wave form. The sound wave is plotted on a horizontal bar to delineate positive (+) from negative (-) air pressure. Moreover, the sound wave is plotted against time, which always moves in one direction (from the past to the future). Here is a very simplified representation of a sound wave:



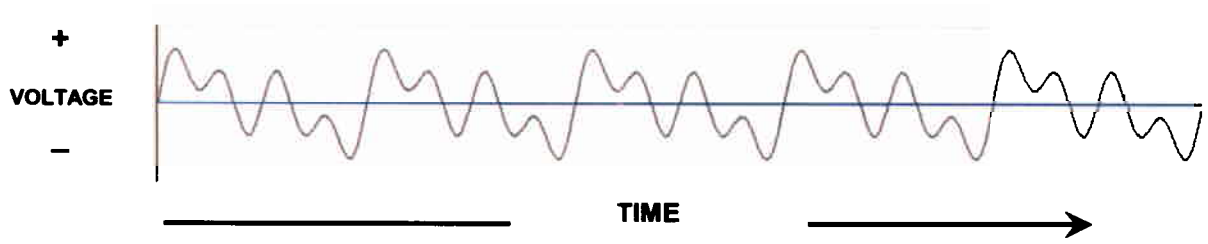
25. Speech recognition systems depend on the physical law that time moves in one direction. That is, these systems break down the sound into small units of time, in sequential order. If time moved back and forth, then a voice recognizer could interpret a word like

"December" as "em-Dec-ber" or similar other nonsensical variation, out of sequence. But because time moves in one direction, the voice recognizer "hears" the "D" sound before "c" sound before the "b" sound.

B. The "Voice Recognizer"

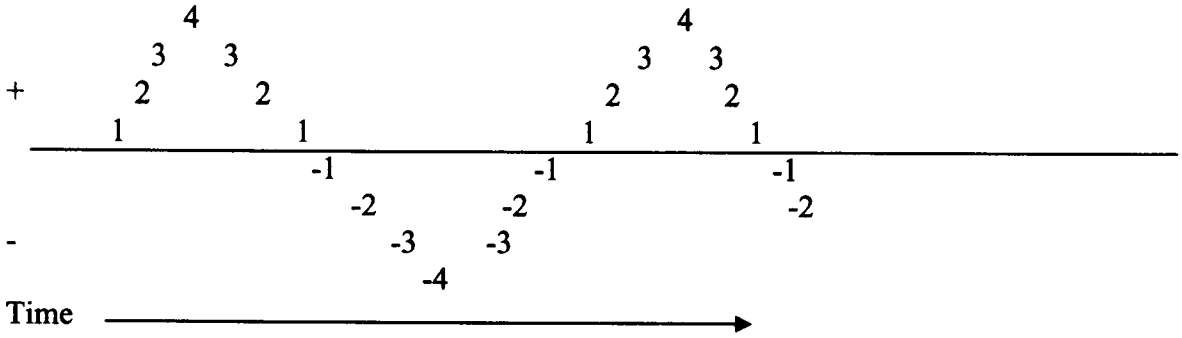
26. There are two basis levels of a speech recognition system. The first is the user interface, which typically consists of prompts or other cues for the user to speak commands. I will discuss this later. The second is known as the "voice recognizer" or "speech recognition engine." Both consist of software running on a computer processor. In the case of cell phones, the software can run on a microprocessor embedded in the circuitry of the phone.

27. In the first step of speech recognition, a microphone, also known as a "transducer," converts the variations in air pressure into corresponding variations in electrical voltage. The resulting electrical signal is known as an "analog" signal because it is analogous to air pressure. That is, the signal is the electrical analog of a sound wave. The analog signal can be diagrammed the same way as the sound wave, except that instead of positive and negative air pressure, we now have positive and negative voltage:



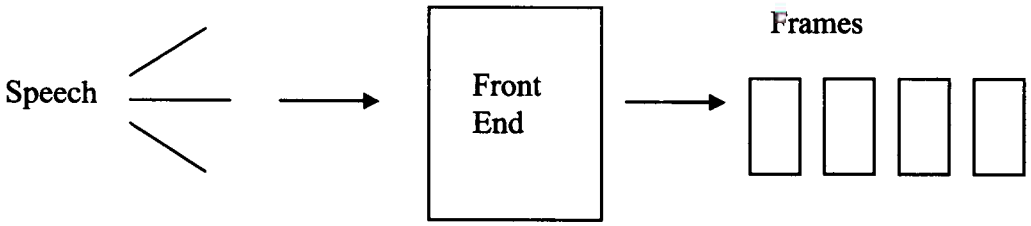
28. After converting the sound to an analog signal, the speech recognition system sends the signal through a digital signal processor ("DSP"), also known as the "Front End" of the

voice recognizer. The Front End converts the analog signal to a digital signal--that is, to a stream of numbers representing points on the sound wave, such as represented below:



B.1. The "Front End"

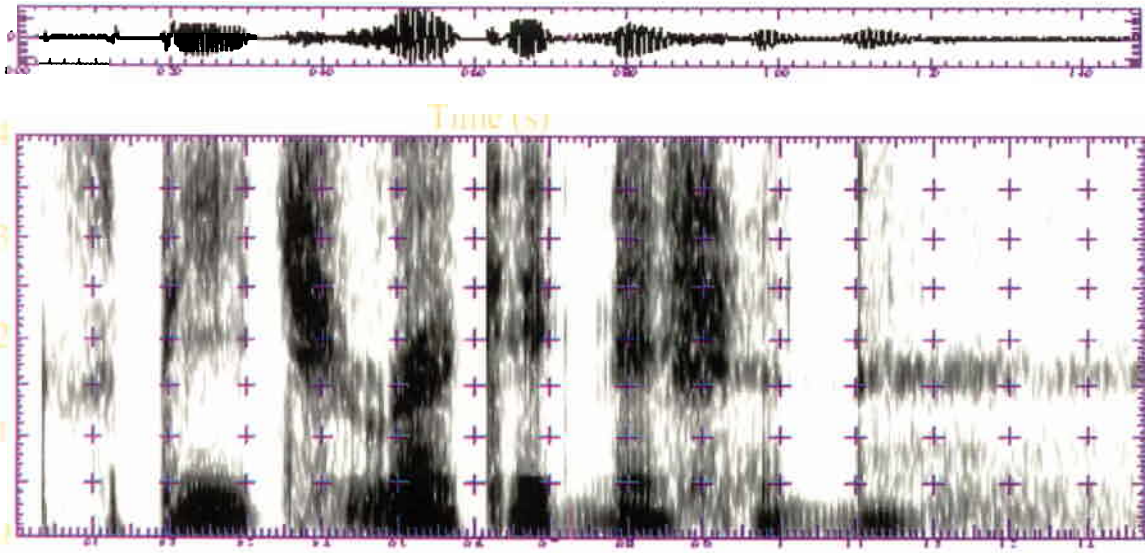
29. The Front End processor converts the raw audio (the sounds produced by the speaker) into a digital stream and then into "frames" comprising "speech vectors." A frame is like a snapshot of sound in any point in time--analogous to individual frames of a movie film. Each frame is a snapshot of 10 milliseconds of time. There are 100 frames of sound per second. Each frame contains numeric values representing measurements--called "features"--which are further abstractions of the raw audio data (e.g., frequency, amplitude, and other acoustical properties of sound). Here is a simplified representation of the Front End processor:



30. Each frame is a snapshot of the spectral quality of sound at any point in time. Each frame represents a slice of the sound spectrum. By analogy, the Front End processor is like

a prism that takes a beam of white light and breaks it out into the colors of the light spectrum. But instead of colors, one gets tones representing the acoustical patterns of speech.

31. Here is an example of a realistic representation of a sound wave. The top box is a "raw waveform" that plots amplitude versus time (as in paragraph 27 above). The bottom box is a spectrogram, a display that shows the sound spectrum of the raw waveform. A sound spectrum can be thought of as energy per frequency band.



32. Of course, the Front End processing is more complicated and typically involves complex "number crunching" to digitize the analog signal, segment it into frames, and extract useful parameters. Powerful algorithms in the Front End software determine the start and end points of speech and perform myriad other functions to process the incoming speech, characterize it, and assign numeric values to its properties.

B.2. The "Back End"

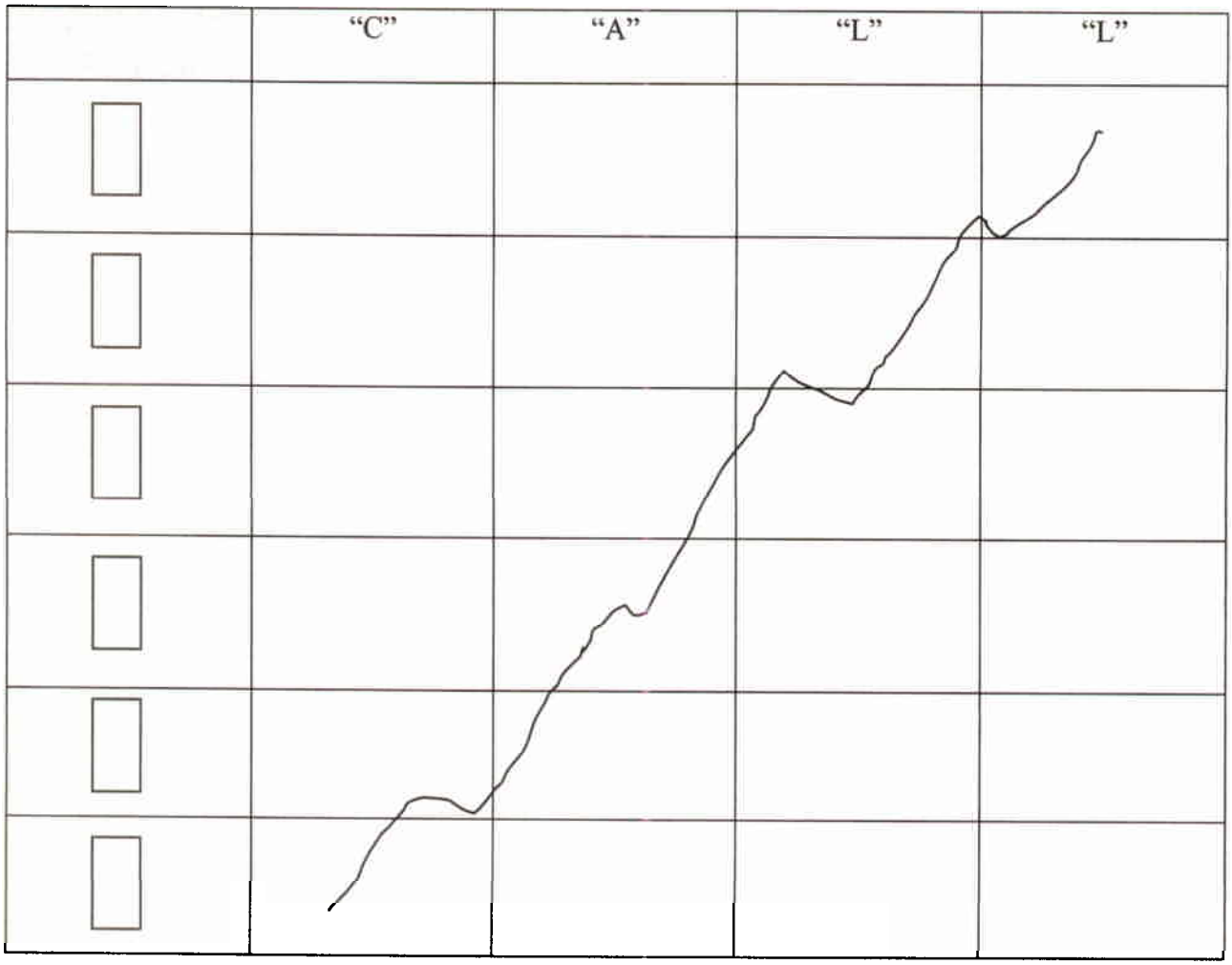
33. After the Front End has processed the raw audio, converting it to digital streams of numbers and then segmenting that stream into frames, the data can be analyzed by the "Back End" of the speech recognition system. The Back End is where the "recognition" happens--i.e.,

where the data processed by the Front End can be analyzed and compared to a vocabulary programmed into the system to determine what words were spoken. The two most common techniques for recognizing speech are (1) Template Matching (also known as "Dynamic Programming" or "Dynamic Time Warping") and (2) Hidden Markov Model statistical analysis.

34. A Template Matching system uses a table or database of previously "trained" (i.e., stored) samples of words and numbers (e.g., the word "call" or the digits 0-9) that users of the system will speak to voice dial a phone number.[1] The system compares the spoken words and numbers with those samples stored in the table.

35. A Template Matching system can be pictured as follows. (In reality, this matching is all done by software, but one can imagine that the process looks like the following table.) Take, for example, a word commonly spoken in VAD systems used to voice dial a phone number: "call." The Front End will segment the sound representing the word "call" into frames. The frames are then compared with sample frames for that word that have been trained into the system. The system them compares each frame with the previously stored sounds. In the table below, the frames representing the word spoken by the user appear in the left column, while the "trained" frames run across the top. The diagonal line represents the "score" for each frame, i.e., the probability or "confidence" level that the frame matches a trained sound.

[1] The system is "trained" as follows: The system developer records hundreds or even thousands of samples of each spoken word or number that the system needs to recognize. The more samples of each word spoken by different people, the more accurate the system will be. The developer in the laboratory can then build the system's vocabulary with these pre-recorded examples. "Training" also refers to a process by which an individual trains the system to recognize his or her particular spoken words. But for now, I am using the term to refer to the process by which the manufacturer of the system builds the table of sample words and numbers.



36. The Template Matching system is highly accurate for recognizing "discrete" or "isolated word speech," which refers to a system that can only recognize spoken words or characters that are surrounded by distinct pauses. That is, the speaker must first pause before and after each word or number.

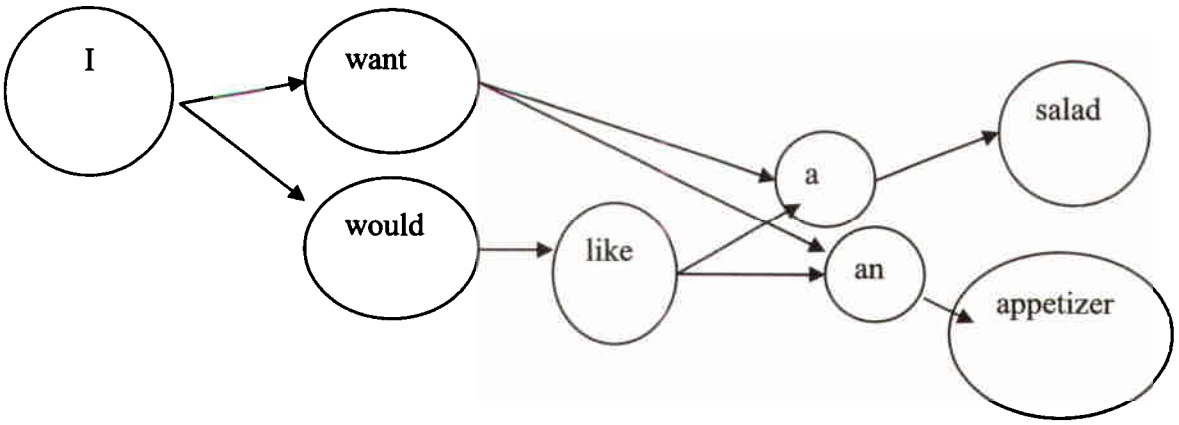
37. A second type of Back End system for recognizing speech is known as the hidden Markov model statistical approach. Markov was a Russian mathematician who devised a method for determining statistical probabilities. Hidden Markov Models are not unique to speech recognition but rather have many uses for predicting events based on segmenting the world into "states." Speech recognition scientists and engineers, however, have adapted this

method of statistical modeling for speech recognition systems. There are several variations of Hidden Markov Models, but for purposes of this tutorial, I will discuss the general concept.

38. Hidden Markov Models segment the universe into "states." In speech recognition systems, these states are defined as basic units of speech. This method allows for the recognition of "continuous" speech--i.e., without the need for pauses between the words or numbers spoken by the system user.

39. In time-dependent systems like speech recognition systems, the next state depends on the previous. Thus, if the first phoneme spoken is a "B" sound, the next phoneme will probably not be "T" because there is no "BT" sound in English. Rather, the statistical modeling allows the system to predict that the next sound will be, for example, an "R" or a vowel.

40. Hidden Markov Models also allow the prediction of the likely next word in the "grammar" spoken by the user. (A "grammar" in speech recognition system is the recognized ordering of vocabulary words needed to direct the system.) For example, in a hypothetical speech recognition system for ordering lunch over the telephone, an appropriate Hidden Markov Model allows the following predictions based on the previous word spoken:



41. Each circle is a "state." The first state helps to determine the next. Speech recognition systems use this style of statistical analysis, albeit in a far more complex way. In

particular, Hidden Markov Model systems employ phonetic recognition. That is, whole words are not predicted and thus recognized, as in the lunch order example above. Rather, phonemes are recognized. But the basic concept is similar.

42. The probability that one state will be followed by another state is based on a statistical algorithm that depends on the context of users, the target application, and other knowledge that is external to the engine. These probabilities are thus "trained" in the laboratory with sample data. The quality of training data determines the accuracy of the subsequent recognition. This modeling process is a method for simulating the all-important context that I mentioned in a paragraph 21 above.

C. Speaker-Dependent and Speaker-Independent Systems

43. There are two basic types of speech recognition: speaker-dependent and speaker-independent. A speaker-dependent system requires the user to "train" the system to recognize words--typically by repeating them several times to provide the system with a useful sample. Speaker-independent systems do not require individual training and allow use by many different people. These systems rely on word and number samples collected from hundreds or thousands of speakers, which are then averaged to build the system's stock vocabulary.

44. A speaker-dependent system is highly accurate for the given speaker because it does not need to recognize so much variation in accents, dialect, and the like. Moreover, this system allows for personalization of commands, such as names. For example, a particular VAD product might allow the user to train the system to recognize the name "John." By uttering that name, the user can command the system to dial John's phone-number.

45. On the other hand, speaker-dependent systems require the user to spend time training the system. As such, these systems are less convenient than speaker-independent systems, which can be used immediately, without training.

46. In contrast, speaker-independent systems are more robust in that they tolerate greater degrees of variation due to dialects, accents, and the like. For illustration only, a speaker-independent system might be designed for specific geographic regions. The system used in, say, the South, may rely on speech samples from Alabama and Mississippi, while a New England-based system might draw samples from, *e.g.*, Down East Maine and Brahmin Boston. Thus, a hypothetical New England-based system will recognize that one has "pahked" a "cah" in "Hahvahd yahd," while the Southern system will recognize "y'all."

47. Of course, modern systems are based on samples of speech from a broad array of regions and are therefore more tolerant of regional dialectal variation. So it would not be necessary, for example, to literally field a different vocabulary for Mississippi and Boston. But the example is illustrative of the problem of "tuning" a vocabulary such that it is most tolerant of its specific users.

D. The User Interface

48. The other main level of a speech recognition system is the user interface software, which manages the dialog between the user and the voice recognizer.

49. The user interface is where my specialty, "human factors" analysis, comes into play. As I mentioned, the basic problem of speech recognition systems is error rate--*e.g.*, recognizing the command "call home" when the user has actually said "call office." Technology developers have focused on this type of error—known as a "substitution" error—because it can be finely measured in a laboratory setting using reproducible experiments and

databases of speech samples. But there are other errors, more insidious, that occur as a result of interactions between the user and the system dynamically.

50. For example, a user may speak the command "call home." If the second half of the command is spoken in certain ways, the recognizer might not catch the word "home," and would therefore not know it was spoken. The machine would therefore prompt for the name, not knowing that the user has already spoken it. The user, confused by the machine's follow-on request, "Who do you want to call?" repeats the word "home" but this time more loudly and slowly—as if speaking to a person who is hard of hearing. This exaggerated speech is now incorrectly recognized—perhaps as "James"—because of the increased variation. One error has caused another.

51. Human factors specialists observe these so-called "error-amplification" effects all the time. In the above example, the substitution error "Home => James" was caused by the change in the user's mental context as a result of a previous error. Speech technologists in the lab do not observe such effects because their databases do not change as a result of errors. So these kinds of human-interaction effects can only be observed in live interactions.

52. We in the speech recognition industry have found two basic methods for reducing and eliminating errors. One way is to go back to the lab and improve the technology used in the speech recognition engine--such as better statistical models, better front end processors, and the like. The second method is to make the application smarter--that is, by using external intelligence beyond the basic acoustic recognition capabilities of the system to prevent or correct errors. That is where the user interface comes in. For example, the error rate can be reduced by making the user prompts and commands less complicated, more intuitive, and otherwise easier to navigate. A smarter user interface helps to prevent errors before they occur (because the user has

spoken the correct command when prompted) or to correct errors made by the voice recognizer (because, e.g., the user interface prompts the user to repeat the command).

53. The construction of an "*n*-best" list is a typical method for preventing or correcting errors. The speech recognition system uses the *n*-best list to rank recognition results. The best match (n=1) is probably the correct result. But if the system is unsure (e.g., due to faulty recognition, distorted speech, etc.), it can offer the user the second-ranked result, and so on, as part of an error-correction dialogue. For example, the user interface might ask the user, "Did you say 'call James?" and allow the user to verify the response. If the user answers, "no," then the application might drop to the second choice, saying, "Then did you say 'call home'?" thereby correcting the error without having to capture new input.

SPEECH RECOGNITION SYSTEMS CIRCA 1992

54. The '966 patent derives from an application filed in 1992. In the late 1980s through 1992, the speech recognition industry was still relatively young. There had been tremendous advances made in the laboratory to develop the basic technology and statistical models used in the Front and Back ends of the speech recognition engines. But there had been relatively little productization of this technology. That is, much of the technology was developed in academia or in corporate R&D labs to demonstrate the concept of speech recognition. But little of it found its way into commercial products. Thus, there was not a lot of work being done on applications of speech recognition technology--on user interfaces, for example. Of course, there were some VAD products in those years, but they were mostly characterized by poor user interface design and thus poor user acceptance. As a result, these products were not very successful and did not gain wide market acceptance.

55. In the 1980s and early 1990s, the cell phone market was also a fledgling industry. In the early 1980s, shortly after the break-up of AT&T, many young telephone network carriers, like Sprint and MCI, began to emerge and began to offer new products, like mobile cellular calling. As I mentioned above, it soon became obvious that hands-free dialing was needed for car phone users. Thus, the speech recognition industry began to address this need and began to develop VAD systems, first at the experimental and academic level, and then later at the commercial product level.

56. VAD products at that time were not very intelligent. That is, the user prompts and commands were more laborious to use, the dialogs less natural and user-friendly, and the system was more error prone. For example, if a user were speaking a telephone number to be dialed, he or she would have to prompt the system to let it know that the user had finished speaking the phone number. The system was not smart enough to recognize when a complete phone number was spoken or to prompt the user to repeat the number if a complete telephone number was not spoken.

57. By analogy, I can say to a friend holding a phone, "dial 617-443-9292," and my friend knows that I have spoken a complete number (and also knows to add a "1" prefix, if necessary). I do not have to say, "OK, that's it. I've finished speaking the number," or "don't forget to dial '1' first." My friend does not have to ask, "what is the next digit?" But VAD systems at the time did require some additional prompt or command to signal that a complete phone number has been spoken. And these systems could not distinguish between digit strings of valid or invalid (i.e., complete or partial) lengths. They would not know to ask the user to correct an incomplete number and would simply attempt to dial whatever digit string the user provided after the user indicated that he or she had spoken the final digit in the string.

58. Conventional, land-line telephone companies, however, have for years relied on algorithms programmed into the networks that look for the last expected digit in the number to be dialed. This system intelligence would know whether to expect, say, a seven-digit or eleven-digit number based on the first number or numbers dialed. That is, even users of old fashioned rotary phones simply had to dial the digits of a phone number and did not have to signal the phone network that the user had finished dialing. The system would determine when all of the required digits had been dialed and route the call accordingly But such system intelligence--what I will refer to as "smart dialing"--had not been built into VAD systems for mobile phone networks,

59. Convenience, economics, and marketing decisions also influenced the form that speech recognition products took in the early days. For instance, cell phones distributed to customers at the time often did not already have the speech recognition circuit boards and software built in. It would have been more convenient and easier to introduce voice activated dialing into the market if the cell phone owners did not have to trade in their phones for upgrades. From an economics point of view, it would initially be cheaper for the network carrier to install the equipment at the switch than to invest in and market upgraded cell phones.

60. Even so, we in the industry understood that it would also be possible and even desirable to place the voice recognizer in the cell phones. One advantage to placing the voice recognizer in cell phones would be to reduce any distortion of the spoken commands over the airwaves from the cell phone user to the recognizer at the switching office. In other words, placing the speech recognition circuitry and software in the phone itself reduces the distance between speaker and recognizer, avoids the "noise" involved in transmission, and thus reduces the chance for distortion.

61. Moreover, while the cost of computing power and memory was relatively high at the time, given Moore's Law (that the number of transistors on an integrated circuit would double every 18 months), that cost was coming down rapidly. The power and speed of microprocessors were increasing dramatically, and the cost was coming down to commodity levels. Thus, even in the late 1980s, some speech recognition engines were being built into the cell phones themselves. I know that the research and experimentation in 1992 and before was not limited to designing switch-based voice recognition systems. For example, the Uniden, Pawate, and Ishii references, all cited in the '966 patent, appear to show the use of speech recognition engines (*i.e.,* the circuit boards and software) embedded in the mobile phones.

THE '966 PATENT

62. I have reviewed the '966 patent and its prosecution history. As stated above, the patent concerns certain speech recognition systems for use in a "mobile telecommunications system," which the patent defines as "cellular, satelite [*sic*] and personal communications network environments." '966 Patent at col. 3, *ll.* 46-48. In other words, various types of "non-wireline" communications networks (meaning communications over airwaves or media other than the land lines of a standard telephone system). Today we use the word "wireless."

63. A cellular network environment includes all components of the network, from the mobile phone unit itself (e.g., a car or cell phone), to the tower (which contains the transceiver for receiving the signals from the cell phone and transmitting them to the switch), to the switch itself. A "switch," also known as an MTX (mobile telephone exchange), is the centralized computer equipment that routes phone calls to their intended destinations. A typical cell phone network may actually have many switches.

64. A mobile telecommunications system can also include a satellite network and a personal communications network. A personal communications network, or "PCS," was a buzz term in the early 1990s, as the telecommunications industry began to expand rapidly. A PCS could be any network or system of communication, including personal computers and even walkie-talkies. .

65. The patent states that speaker-independent and speaker-dependent subsystems can be combined in the patented speech recognition system. According to the patent specification, the speaker-independent subsystem "allows the user to interact with the system employing non-user specific functions." Likewise, "[u]ser specific functions are controlled with the speaker-dependent recognition subsystem 25. User specific attributes collected by the recognition subsystems are stored in the data storage subsystem" '966 Patent at col. 4, *ll.* 21-37. In other words, the speaker-dependent subsystem can be used to recall pre-stored telephone numbers, perhaps based on keywords specific to a user, like names of the user's friends and family. The Speaker-independent subsystem can be used to voice dial phone numbers simply by reciting the number (or by reciting generic keywords like "home" or "office") without having to first train the system to recognize the generic keywords or numbers spoken by the individual user. As such, the system is ready to use, out of the box, without training (except for any personal keywords that the user might wish to store).

66. I understand that ScanSoft alleges that Voice Signal Technologies ("VST") infringes Claims 1-6 of the '966 patent. Claim 1 reads as follows:

> 1. A speech recognition method for a mobile telecommunications system which includes a voice recognizer capable of recognizing commands and characters received from a mobile telecommunications user, the method comprising the steps of:
>
> receiving a command from the mobile telecommunications user;
>
> determining whether the command is a first or second command type;
>
> if the command is the first command type, collecting digits representing a telephone number to be dialed received from the mobile telecommunications user; and
>
> if the command is the second command type, determining whether a previously stored telephone number is associated with a keyword received from the mobile telecommunications user.

67. To one of ordinary skill in the art, this claim refers to an application of a speech recognition system for use in a mobile telecommunication system. The preamble of the claim states that the system includes a voice recognizer. A "voice recognizer," as mentioned above, comprises the software and/or hardware (e.g., a microprocessor) used to process speech, as described above.

68. The speech recognition method includes the steps of receiving a command from a mobile telecommunication user, determining whether the command is a first or second command type, and performing additional steps depending on the command type.

69. The first step, "receiving a command," is self-explanatory. The user (e.g., the user of a cell phone, PDA, or other wireless communications device) says a "command," which is merely a word or number (or combination of words and/or numbers) that the speech recognition system recognizes and that is used to direct the system to take some action. The speech recognition system merely receives the command.

70. The speech recognition system then determines whether the command is a first or second command type. These are merely two categories of commands recognized by the speech recognition system. For example, one category of commands might be various commands used to voice dial a phone number by merely speaking the digits. A second category might include keyword dialing (e.g., "Call home" or "John"). A "keyword" is simply any word or number (or combinations of words and/or numbers) used to recall a phone number that the user has previously stored in the system.

71. In this case, when the first type of command has been spoken, the speech recognition system collects "digits representing a telephone number to be dialed." For example, the user says the command, "Call 1-617-443-9292," and the system (a) determines that it is a first command type and (b) collects the phone number portion of the command so that it can transmit it to the telephone network. The system is intelligent in that it is programmed to know that the digit string spoken by the user must comprise an expected number of digits corresponding to the length of a valid, complete telephone number. I have previously referred to this concept as "smart dialing." There may be various ways for the system to determine when a complete telephone number has been spoken. The patent suggests one method in which the intelligent system is programmed to look for the last digit in the expected string. For example, if the first digit spoken is "4," then the system knows to expect three digits in all -- i.e., "411." Indeed, phone companies typically use algorithms to determine whether to expect, say, 7 or 11 digits when a phone number is dialed by hand. So this intelligence can be programmed into the speech recognition system through various algorithms.

72. The word "representing" in Claim 1 is key to this concept of determining whether the digit string spoken comprises the correct number of digits for a valid, complete

phone number. The word suggests to me that the system does not just collect any number of digits but rather digits that must "represent" or comprise a telephone number. A "dumb" system (i.e., one without the necessary level of built-in intelligence) would simply collect digits forever until told to stop and would have no expectation that the number of digits spoken must represent a valid, complete phone number. A dumb system would simply attempt to dial whatever digits were spoken by the user. In contrast, the speech recognition methods of Claim 1 will not dial or collect a spoken digit string if that digit string is comprised of an unexpected number of digits.

73. When the user speaks a second category of command associated with "keyword" dialing, the system checks to see whether the keyword spoken by the user is already in the system and has a previously-stored telephone number associated with it.

74. I understand that VST has challenged the interpretation of some of these terms in Claim 1. In particular, I understand that VST contends that the voice recognizer must be located at the switch and has pointed to the wording "A speech recognition method for a mobile telecommunications system which includes a voice recognizer . . .". I fail to see how this wording requires that the voice recognizer reside at any particular location. A speech recognition method for a mobile telecommunications system is simply one that is especially adapted for use in the cellular or other wireless network environment (as opposed to, say, a speech recognition method used in a call center or to control one's personal computer). The speech recognition hardware and software perform the method, not the network itself.

75. Indeed, nothing in Claim 1 requires that the voice recognizer reside in any one place. The voice recognizer could be anywhere in the mobile communications system, which by the patent's own definition includes the cell phone or other communications device. In contrast, some of the claims of the earlier patents in the family specify that the voice recognizer must be at

the switch. For example, Claim 1 of U.S. Patent No. 5,297,183 (the first patent to issue based on the original 1992 application) specifies that "a voice recognition system located at the mobile telecommunications switching office." No such words appear in Claim 1 of the '966 patent, however. The exact location of the speech recognition system is not specified. It could reside anywhere in the network environment.

76. Furthermore, as I stated in my discussion of the state of the art circa 1992, there was no particular technical reason for placing the voice recognizer at or near the telephone exchange switch. The industry realized back then that one could put the recognition system in the mobile phone, in the trunk of a car with a mobile car phone, or anywhere else within the network environment.

77. The patent specification states that placing the speech recognition system at the switch "reduces costs and increases reliability by enabling the switch vendor to install and maintain the system in conjunction with the cellular switch." '966 Patent at col. 1, *ll.* 51-54. That reasoning is what I discussed above--economics and marketing decisions might influence placing the speech recognition equipment at the switch when first introducing a VAD system to the market, but technology constraints did not require it. Thus, the patent also teaches that "[m]any other beneficial results can be attained by modifying the invention as will be described," col. 2 at ll. 29-30, and that placing the speech recognition system at the switch was preferable at the time but that "those skilled in the art" could use the preferred embodiments disclosed in the specification as a starting point, "as a basis for modifying or designing other structures for carrying out the same purposes of the present invention." Indeed, early products—based on "dumb dialing" and with extremely limited capabilities—could be found in products such as the Uniden VoiceDial product. It is well within the ordinary skill of those in the art, having read the

'966 patent, to implement the voice dialing methods of the claims anywhere within the network environment, including the switch or the handset of the mobile phone.

78. It is clear to me from reading the patent, and from what I know about the evolution of speech recognition in cell phone networks, that one of the purposes of the speech recognition methods of Claims 1-6 was to improve and make more robust a hands-free, voice activated dialing cell phone by improving the speech recognition system's user-interface (e.g., the command and prompt sequences) and making the system more intelligent (by using "smart-dialing," for example). In other words, using "human factors" to improve the system, rather than advances in microphones or integrated circuits. My review of the prosecution history shows that placing the voice recognizer at the switch was not important to the speech recognition methods claimed in the '966 patent--i.e., that placement was not used to distinguish the prior art.

79. VST further appears to contend that the step of "determining whether a command is a first or second command type" requires some separation between a "command" and the series of digits or keyword that the user would speak next. I do not see this distinction. A valid command could be, simply, "Call Home" or "Dial 1-617-443-9292". In other words, the command could be "call" or could be the combination of "call" and a keyword, or simply the keyword (or string of digits representing a telephone number) itself.

80. Claim 6 of the patent confirms my analysis. Claim 6 provides that the method further comprises "the step of prompting the mobile telecommunications user to enter information needed for the first or the second command type." In other words, as seen in the preferred embodiments of the patent (illustrated in the flow charts), a preferred method would be to have the user say "call," and the system would respond, "state a number" or "state a

keyword." Claim 6 appears to cover this method, which must mean that Claim 1 covers that method and others as well (such as having the command include the number or keyword).

81. In summary, I believe that one of ordinary skill would have known that variations on the invention could be built and that nothing in the patent constrains the system as VST interprets it.

I DECLARE UNDER PENALTY OF PERJURY THAT THE FOREGOING IS TRUE AND CORRECT. EXECUTED ON MAY 6, 2005.

/s/ Bruce Balentine
BRUCE BALENTINE

02639/00509 379105.1

keyword." Claim 6 appears to cover this method, which must mean that Claim 1 covers that method and others as well (such as having the command include the number or keyword).

81. In summary, I believe that one of ordinary skill would have known that variations on the invention could be built and that nothing in the patent constrains the system as VST interprets it.

I DECLARE UNDER PENALTY OF PERJURY THAT THE FOREGOING IS TRUE AND CORRECT. EXECUTED ON MAY 6, 2005.

BRUCE BALENTINE

02639/00509 379105.1